DOCUMENT RESUME

ED 305 648 CS 211 761

AUTHOR Baldwin, Janet

TITLE Writing Skills of Graduating High School Seniors and Adult High School Non-Completers: A Study of Factor Structure Invariance.

PUB DATE 28 Mar 89

NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 28, 1989).

PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Adult Students; Dropouts; High School Equivalency Programs; High Schools; High School Seniors; Models; *Structural Analysis (Linguistics); Writing Instruction; Writing Research; Writing Skills

IDENTIFIERS *Confirmatory Factor Analysis; Factor Invariance; *General Educational Development Tests; Linear Relationships; Writing Skills Assessment Test

ABSTRACT

A study examined factor structure invariance among the writing skills of graduating High School Seniors and Adult High School Non-Completers. The study had three purposes: (1) to use LISREL confirmatory factor analysis (CFA) procedures to specify and test a series of factor models based on the test specifications content of the General Educational Development (GED) Writing Skills Test; (2) to evaluate the fit of these models to a set of test data obtained from a national sample of graduating high school seniors; and (3) to test the invariance of the best fitting factor structure model in both the seniors and the adult high school non-completers. Subjects were 2,532 high school seniors who took the anchor form of the Writing Skills Test in the spring of 1987 and 699 adult high school non-completers between the ages of 17 and 19. Results indicated the nature of writing skill measured by this test to be a single construct reflecting generalized proofreading/editing skills. No support was found for the view that separate item-type methods factors accounted for variability of performance on the multiple-choice portion of the GED test. This study provided an empirical field-based illustration of the use of CFA procedures to evaluate the factor structure of multiple-trait, multiple-method data and to test for the invariance of plausible measurement models over multiple groups. (Two tables of data are included, and 24 references and a list of variables are attached.) (RAE)

Writing Skills of Graduating High School Seniors and Adult High School Non-Complcters: A Study of Factor Structure Invariance

Janet Baldwin

American Council on Education, Washington, D.C.

Presented at the Annual Meeting of the National Council on Measurement in Education in San Francisco on March 28, 1989.

Writing Skills of Graduating High School Seniors and Adult High School Non-Completers: A Study of Factor Structure Invariance

by

Janet Baldwin
American Council on Education

Various theories about the nature of writing skill address such questions as whether writing skill is a single generalizable construct or comprised of discrete multiple skills; whether writing performance is influenced by the nature of the task, assignment, or question format which elicits such performance; or whether writing skills are context based, influenced by the various subject matter domains within which writing may be assessed. Because some writers have argued that "there is always more than one major factor underlying any test data in the achievement domain. . . . even when achievement in a specific topic is being measured" (Birenbaum & Tatsuoka, 1982; p. 259), theories about test data in the domain of writing achievement also may address the multidimensionality of writing achievement test data. Because the Tests of General Educational Development are designed to measure the major and lasting outcomes of a four year high school program of study, the GED Writing Skills Test is intended to assess generalized writing skills which correspond to what graduating high school seniors are required to know and to do. Although it may be reasonable to expect a test of generalized knowledge and skills to be unidimensional (Phillips & Mehrens, 1987), assumptions about what the test measures should also be examined empirically. Performance on a multiple-choice writing test which includes items in multiple content or skill areas should be evaluated to determine whether the test actually measures distinct multiple skills or, rather, whether it indicates a more generalized set of editing or proofreading capabilities. Examining what the test measures in the various populations for which it is intended not only improves the interpretability of the test scores but increases knowledge about the nature of writing skills in different groups.

Although the test specifications for the GED Writing Skills Test are used primarily as a blueprint for test assembly, the classification of items in the test specifications may also be used to provide a conceptual basis for evaluating the nature of what the test actually measures. For example, item type classfications from the test specifications of the SAT-V and SAT-M subtests have been used as a basis for examining the construct validity of the SAT across populations (Rock & Werts, 1979). In the current study the test specifications were used as a basis for defining both writing skill and item type factors in order to examine the internal construct validity of the GED Writing Skills Test (Part I) across populations.

This study had three purposes. The first purpose was to use LISREL confirmatory factor analysis (CFA) procedures (Joreskog, 1979; Joreskog & Sorbom, 1983; Long, 1983) to specify and test a series of factor models based on the test specifications content of the General Educational Development (GED) Writing Skills Test. The writing test content included three writing skills, or traits, and used three multiple-choice item-types, or methods, of measurement. The second purpose was to evaluate the fit of these models to a set of test data obtained from a national sample of graduating high school seniors. The third purpose was to test the invariance of the best fitting factor structure model in both the seniors and the adult high school non-completers.

Instrument and Data Source

The GED Writing Skills Test is one of five tests in the GED test battery which were developed to measure the major and lasting outcomes of the learning associated with a high school education. Although the high school equivalency tests are administered to a diverse group of adults who did not complete a typical four-year high school program of study, these tests are standardized on a nationally representative sample of graduating high school seniors in order to provide the score scale and the basis for passing requirements. The GED Writing Skills Test has recently been revised to reflect changes in high school writing curricula which place more emphasis on essay writing and proofreading skills than in the past (GED Testing Service, 1987). In order to facilitate the interpretation of test scores in these groups, it is important to know whether the test measures the same writing constructs in both adult high school non-completers, for whom the test is intended, and graduating high school seniors, on whom the test is standardized.

The GED Writing Skills Test has two parts. Part I contains 50 multiple-choice items[1] and Part II consists of a single essay topic. Data used in the analyses for this study are limited to multiple-choice items from the anchor form of the test. The multiple-choice portion of the test measures the ability to edit and correct errors within the context of one or more paragraphs of extended discourse. The basis for

[1] In addition to 50 multiple-choice items which are scored, Part I of the Writing Skills Test also contains 6 non-scored field test items.

the items is a number of selections of prose, each consisting of one or more paragraphs. Errors in sentence structure, usage, and mechanics occur throughout the selection and reflect the types of errors that examinees typically make in their writing. These errors are tested in the items that follow each passage.

In order to minimize problems associated with factoring dichotomous data (Gorsuch, 1983), fifteen mini-scale variables were created by summing across dichotomous test items of similar content. All multiple-choice items were classified by writing specialists according to type of editing skill measured and item type. Each miniscale variable was based on 3 to 5 test items which had been judged to measure the same skill (trait), using the same item type (method) (Attachment A).

The test was administered to a nationally representative sample of high school seniors in the spring of 1987 as part of a national standardization study conducted by the test developer. For purposes of these analyses, only those graduating high school seniors who took the anchor form of the Writing Skills Test (N = 2,532) were included. From January through June of 1988, the test was administered to adult GED examinees seeking to qualify for a high school equivalency diploma. Because the influence of age on factor structure was not the specific focus of this study, it was necessary to control this variable in the adult GED examinee sample where ages ranged from 16 to 87. Therefore, only those adult high school non-completers between the ages 17 and 19 (N=699) were included in the "young adult" sample for these analyses. This age range was selected because it is nearest to the age range of graduating high school seniors. For each group, covariance matrices of measured variables were used as input data for CFA procedures.

Method

Preliminary exploratory factor analyses applied both principal factor and maximum likelihood procedures to the dichotomous item level data in order to determine whether the items clustered together in a way which supported the miniscale item/variable groupings. Second, a series of factor models were specified to represent test content. These models specified up to three traits -- proofreading and editing in the areas of 1) sentence structure, 2) usage, and 3) mechanics -- and up to three methods, or item types -- 1) sentence correction, 2) sentence revision, and 3) construction shift -- according to content categories in the test specifications table of the multiple-choice portion of the GED Writing Skills Test (GED Testing Service, 1984; 1987). Confirmatory factor analyses were applied to the miniscale level data from the national sample of graduating high school seniors (N = 2532) in order to determine the best-fitting model for subsequent multiple-group hypothesis testing. Third, using simultaneous CFA for two groups, the best-fitting model was tested for goodness of fit to data from both the sample of graduating seniors and the sample of adult high school non-completers.

CFA procedures. Eleven models were specified and tested using confirmatory factor analysis procedures. Model 1 was a null model, with no common factors. Model 2 hypothesized a single factor influencing all fifteen variables. Models 3, 4, 5, and 6 were chosen to represent relationships among writing skill content factors. Models 7, 8, 9, and 10 were chosen to represent relationships among the item types factors.

Model 11 represented a factor model which included three writing skills and three item types.

<u>Model 1</u>. Null Model - 15 uncorrelated factors;

<u>Model 2</u>. One Factor Model;

<u>Model 3</u>. Two Correlated Trait Factors (Traits 1 and 2 combined into one factor);

<u>Model 4</u>. Two Correlated Trait Factors (Traits 1 and 3 combined into one factor);

<u>Model 5</u>. Two Correlated Trait Factors (Traits 2 and 3 combined into one factor);

<u>Model 6</u>. Three Correlated Trait Factors;

<u>Model 7</u>. Two Correlated Method Factors (Methods 1 and 2 combined into one factor);

<u>Model 8</u>. Two Correlated Method Factors (Methods 1 and 3 combined into one factor);

<u>Model 9</u>. Two Correlated Method Factors (Methods 2 and 3 combined into one factor);

<u>Model 10</u>. Three Correlated Method Factors;

<u>Model 11</u>. Six Factors: three correlated trait factors and three correlated methods factors.

In these single group confirmatory factor analyses, all latent factors were constrained to have unit variances and factor correlations for all models were freely estimated. Criteria for model-data fit included chi-square difference tests (Joreskog & Sorbom, 1983) and several indices of fit produced as output in the LISREL program: chi-square (probability) value, Goodness of Fit index (GFI), Root Mean Square Residual (RMR), and Normalized Residuals (NR). In addition, the Parsimonious Fit Index

(PFI)[2] (James, Mulaik, and Brett, 1982) and ratios of chi-square to degrees of freedom[3] (Joreskog & Sorbom, 1979) were used. Finally, judgments were made about which of the models provided the most plausible and parsimonious representation of the data.

Because the chi-square value is dependent on sample size, the chi-square probability value in very large samples may be significant even when the model represents the data quite well. In small samples it may be non-significant even for models which are poor. Therefore, in analyses based on large samples, the chi-square probability value can lead to rejection of a good model, thereby reducing its usefulness as an indicator of goodness of fit. Hayduk (1987) noted that chi-square is instructive as an indicator of fit for samples ranging in size from about 50 to 500, although this range may vary depending on the kind of model to be estimated. Samples larger than 500, he observed, require other indices of fit. Criteria of fit which may be more relevant are sequential tests of incremental differences in fit, or chi-square difference tests, because such tests improve inference with both large and small samples (Bentler, 1980). Because the differences in chi-square

---

[2] The parsimonious fit index (PFI) is actually Bentler and Bonnett's (1980) normed fit index modified to take into account the number of degrees of freedom given up in order to arrive at a particular level of goodness of fit. Generally, the models with the maximum values of PFI are those that best describe the data with the fewest unknown parameters (Loehlin, 1987). The formula for the PFI, where o refers to the null model and k refers to the compared model, is:

$$(df_k/df_o) \times \frac{(\text{chi-square}_o - \text{chi-square}_k)}{\text{chi-square}_o}$$

[3] The range for recommended ratios of chi-square/degrees of freedom (df) typically are between 2 and 5 (Carmines & McIver, 1981).

values are themselves chi-square statistics, they can be used to test the importance of parameters that differentiate nested models.

Invariance Analyses. The model judged to provide the best fit to the data from the seniors sample was then tested for equality of factor structure over samples of seniors and young adult GED examinees. In each analysis, latent factor variances were freely estimated and indicator variables were selected to set the metric for each factor. The following invariance hypotheses were tested:

Equal SIGMA. Equal covariance matrices in both groups.

Equal k. The factor analysis model has the same structure in both groups; i.e., the measured variables load on the same number of factors in the same pattern in both groups.

Equal LAMBDA. In addition to the constraints of Equal k, the factors are measured in the same units in both groups; i.e., equal factor loadings.

Equal THETA. In addition to previous constraints, the factors are measured with the same accuracy in both groups; i.e., error variances are the same in both groups.

Equal PHI. In addition to previous constraints, the variances and covarances of the latent factors are equal.

In an analysis of variance framework, whenever the primary purpose of sampling is to make comparisons across subgroups, the optimum sample is one where the sample sizes of the subgroups are equal (Sudman, 1976). Because the sample sizes for seniors and GED examinees were not only very large but also unequal -- 2,532 and 699, respectively, -- it was anticipated that judgments about goodness of fit could be confounded to an unknown extent and that chi-square tests of factor structure invariance may be highly significant regardless of how well the factor model represented the data. In pointing out the limitations of the use of the chi-square value as a goodness of fit indicator in large samples, Hayduk reports Hoelter's (1983) recommendation concerning the use of a "critical-N", the sample size that would be required to make the observed differences between the estimated and the observed covariance matrices just significant at a typical level such as .05. After examining numerous models, Hoelter suggested that a reasonable sample size cut-point for CFA hypothesis testing is a critical-N of 200 or more. However, because problems of nonconvergence and instances of improper solutions have been found for sample sizes less than 400 (Boomsma, 1985), a critical-N of between 400 and 500 may be more appropriate. Hayduk (1987) pointed out that Hoelter's decision criterion can be obtained by simply inserting the critical-N sample size into the LISREL program, using the observed covariance matrix computed on the basis of the actual sample size. This approach, in effect, ignores the extra sensitivity or precision provided by the extra cases in the sample.

In order to consider whether decisions about goodness of fit may be clarified using this approach, invariance analyses were carried out using both the full sample sizes as well as a critical-N of 400 for each group. The critical-N analyses used the original covariance matrices and substituted the smaller sample sizes in the LISREL programs. This approach was applied in order to estimate the chi-square probability values for samples just large enough to detect meaningful differences -- i.e., a sample size for which the chi-square may be considered an instructive indicator of fit (Hayduk, 1987).

Results and Conclusions

Preliminary EFA. Results from both principal factors and maximum likelihood methods supported the view that there was only one non-trivial factor which accounted for common variance in the raw data. The first factor had an eigenvalue of 18.2. While a second factor had an eigenvalue slightly over 1.0, only one of the 50 multiple-choice items had a loading on this factor which was greater than .30 and this loading was only .31. Finding support for only a single factor at the multiple-choice item level suggested that a single factor may also account for relationships among the 15 miniscale variables as well. Nevertheless, for the purpose of completeness, all models which were proposed were tested.

CFA Procedures. Six of the eleven models tested -- Models 4, 5, 8, 9, 10, and 11 -- produced non-positive definite PHI matrices, making interpretation of the results for these models questionable. However, based on the single factor results from the EFA, and on the generalized nature of proofreading and editing skills the multiple-choice portion of the test was designed to measure, it was not surprising that some models with two or more factors proved untenable. For each of the remaining five models, all estimated parameters were highly significant and there were no instances of improper solutions. Goodness-of-fit results for these five models -- the null model (Model 1), the one factor model (Model 2), models with two correlated traits (Model 3 and Model 5) and a model with two correlated methods (Model 7) -- are presented in Table 1.

Although the chi-square results for these five models were highly significant ($p < .0005$), suggesting that the proposed models do not fit the data, the significant chi-square results were expected given the very large sample size for the high school senior group ($N = 2532$).

Therefore, other indices of fit were evaluated in order to select the best fitting model.

Each of the two-factor models -- Model 3, Model 5, and Model 7 -- produced highly similar goodness of fit indices. Chi-square values ranged between 617.10 and 620.65 with 89 degrees of freedom. For each of these three models, the GFI was .962, the RMR was .023, and the number of normalized residuals (NR) greater than 2.0 was 10. The value for the PFI was the same for these three models, .82. Based on these goodness of fit indices, there appeared to be no discernible differences in these two-factor models. The goodness of fit indices for the one-factor model -- GFI, RMR, and the number of NR greater than 2.0 -- were identical to those for the two factor models. Because the chi-square value had an additional degree of freedom (621.54 with 90 df), the PFI was a slightly improved .83 compared to the PFI of .82 for the other models. This suggests that the PFI as a goodness-of-fit index is sensitive to the advantages of increased parsimony in the one-factor model. When the fit indices for a simpler factor structure appear to be nearly the same as those for more complex structures, the more parsimonious model usually provides the better representation of the data.

An examination of the estimated correlation between the two latent factors in each of the two-factor models provides additional support for the one-factor solution. For Model 3, Model 5, and Model 7, the estimated correlation between latent factors was .990, .996, and .981, respectively. When standard errors for these estimates are taken into account, .005, .004, and .009, respectively, the latent factors in each two-factor model are, in effect, perfectly correlated -- implying a single factor. Based on evaluation of all the CFA results, then, Model 2 --the one factor model -- was selected as providing the best, most parsimonious, fit to the data for testing subsequent invariance hypotheses.

Invariance analyses. The results of the tests for invariance of the one-factor model are presented in Table 2. In the first row for each hypothesis are results using the full sample sizes for seniors and examinees (N = 2532 and N = 699, respectively). In the second row for each hypothesis are results based on the critical-N sample sizes (N = 400 for both groups).

When based on the full sample sizes, the chi-square values for all invariance hypotheses were highly significant. Based on the large and unequal sample sizes for the two groups, significant chi-square values were not unexpected. However, the GFI and RMR improved considerably from the Equal-SIGMA to the Equal-k hypothesis (from .951 to .972 for GFI; from .146 to .026 for RMR), suggesting that the one-factor solution provides a satisfactory fit for both groups. For the Equal-k and Equal LAMBDA hypotheses, the GFI results are .972 and .967, respectively, and the RMR results are .026 and .044, respectively. Although these indices suggest a slight decline in fit for the Equal-LAMBDA model, the differences between the fit indices under each hypothesis are very small. Indeed, one could argue that these differences are not substantively important and that these results in fact support the conclusion that factor loadings (LAMBDA) are equal across groups. However, there are no known statistical criteria for evaluating how large such differences should be in order to judge them important. When the chi-square difference test was applied, the result was a statistically significant chi-square difference of 32.19 with 14 degrees of freedom between the Equal-k and the Equal-LAMBDA hypotheses. If based on statistical significance alone, this outcome would indicate that constraining the one factor model to equal factor loadings across groups results in a significantly poorer fit to the data. However, the chi-square difference of 32.19 represents only a 4% decrement in fit for the Equal-LAMBDA model (806.27) compared to the chi-square for Equal-k (774.08), which suggests that these differences may not be meaningful.

In the comparison between Equal-LAMBDA and Equal-THETA, all goodness-of-fit indices suggest a much poorer fit for the Equal-THETA model. The GFI and RMR values become poorer (.951 and .050, respectively), the chi-square difference test is highly significant (129.12 with 15 df), and the chi-square value represents a 16% poorer fit. Therefore, the results based on the large sample sizes provide support for the Equal-LAMBDA model but not for the Equal-THETA model.

If the chi-square test using a large sample size is viewed as providing so much power that trivial differences are magnified in importance, then it may be instructive to determine whether the chi-square (probability) test -- which is not a useful indicator of fit in large samples -- produces an indication of better fit when power (sample size) is reduced so that only substantively meaningful differences are detected. Although this line of reasoning has serious shortcomings[4], let us consider the goodness-of-fit results based on sample sizes of 400.

---

4 This approach assumes that poor models would be rejected at a sample size of 400. However, there is no apparent basis for quantifying substantively meaningful differences a priori in terms of sample size.

For the analyses based on a critical-N of 400, all invariance tests for the one-factor model produced non-significant chi-square values except for the test for an invariant PHI matrix. The non-significant chi-square for SIGMA ($p = .106$) suggests that the matrices of writing test data for seniors and GED examinees are, in effect, equivalent. Equivalence at this stage implies that the relationships among the measured indicators have equivalent psychometric properties across groups. Although such an outcome implies invariance of factor structure, results from subsequent invariance tests are presented for the purpose of completeness and in order to evaluate the specific nature of invariance in these groups. The non-significant chi-square for Equal-k ($p = .380$) supports the conclusion that a one factor model provides a very good fit to both sets of data. In addition, the test of invariant LAMBDA ($p = .390$) suggests that the factor loadings in each group are equivalent, indicating that the observed variables are measuring the same writing skill construct in the same metric in both groups. With a critical-N of 400, the chi-square probability results support the conclusions about invariance of factor loadings (Equal-LAMBDA) which were based on judgments about the large sample GFI, RMR, and chi-square difference results. That is, the chi-square probability values associated with the tests for invariance of both Equal-k and Equal-LAMBDA indicate very good fits to the data in both groups. Although the critical-N test for equality of THETA is non-significant ($p = .054$), suggesting a good fit, it is only just so. A more conservative interpretation would suggest that the measures, while highly similar, may not be equally precise in both groups. No support was found for equality of variances for the writing skill factor, or PHI, in both groups (p = .016). An examination of the values for the latent factor variance in each group suggests that writing skill performance in the senior group, as measured by the multiple-choice writing test, varies nearly 50% more than in the GED examinee group (1.438 and .951, respectively).

Conclusions and Implications.

First, no support was found for the view that separate item-type methods factors accounted for variability of performance on the multiple-choice portion of the GED Writing Skills Test. Second, the nature of writing skill measured by this test appears to be a single construct reflecting generalized proofreading or editing skills. That is, proofreading and editing writing skills which involve sentence structure, usage, and mechanics do not appear to provide systematically distinct sources of variation in performance on this test. Third, the writing achievement measured by the multiple-choice portion of the GED Writing Skills Test has a highly similar factor structure in both high school seniors and young adult GED examinees. That is, the writing test achievement of young adult GED examinees and high school seniors is based on similar types of writing skills and knowledge. Fourth, this study provided an empirical, field-based illustration of the use of CFA procedures to evaluate the factor structure of multiple-trait, multiple-method data and to test for the invariance of plausible measurement models over multiple groups.

Both EFA and CFA procedures suggested that a one-factor model accounted for common variability in the data. CFA tests of two-factor models found estimated correlations between the two latent factors of virtual unity. These empirical outcomes, along with substantive knowledge about the generalized nature of the writing skills which the test was designed to measure lead to the selection of the one factor model as providing the best, most parsimonious fit to the data for the seniors.

The goodness of fit indices for the large sample invariance analyses suggested support for Equal-k and Equal-LAMBDA. The relatively high GFI values (.967 and .972, respectively) and low RMR values (.026 and .044, respectively) indicated that most of the meaningful variation in the data had been accounted for in these models. However, the chi-square difference test indicated that constraining the one-factor model to have equal factor loadings across groups resulted in a signicantly poorer fit to the data, producing some ambiguity in the decision to retain the LAMBDA hypothesis. Because the chi-square for Equal-LAMBDA represented only a 4% decrement in goodness-of-fit, the chi-square difference was judged not to be meaningful.

A re-analysis of the invariance hypotheses using just large enough, but not too large, sample sizes produced chi-square values for the Equal-SIGMA, Equal-k, and Equal-LAMBDA models which were highly non-significant. If it is reasonable to assume that samples with a critical-N of 400 have sufficient power to detect meaningful differences in the factor structure between the two groups, but not trivial ones, then the results based on the critical-N analyses support the finding of factor structure invariance across groups of both seniors and young adult GED examinees. Because the critical-N invariance analyses were applied to the same data on which previous invariance analyses were based, however, it cannot be interpreted as providing a true test of factor structure invariance. It is recommended that the one-factor model be cross-validated using new sets of data for seniors and young adult GED examinees, with sample sizes of at least 400 but no more than 500, in order to provide an independent test of the invariance of the one-factor model. In addition, future research should address the relationship between sample sizes based on the concept of a critical-N and judgments of what constitutes meaningful differences in model-data fit.

Because these invariance analyses included only those GED examinees similar in age to the high school seniors, it is recommended that similar studies be undertaken to determine if the one-factor model provides a good fit to data for older GED examinees as well. Future studies should also include direct measures writing skill (essay) in order to determine the relationship between both indirect and direct measures of writing performance in seniors and GED examinees.

Although MTMM data have been examined using the Campbell-Fiske criteria (Campbell & Fiske, 1959) and an ANOVA model (Kavanagh, McKinney & Wolins, 1971; Stanley, 1961), CFA models have been found to provide better tests of these matrices without the limitations inherent in the other approaches (Marsh & Hocevar, 1983; Werts, Joreskog, & Linn, 1972). This empirical study of writing test data from high school seniors and young adult high school non-completers not only illustrated the application of a useful methodology for examining MTMM data, but it also contributed to our understanding of writing skill, a construct which is becoming increasingly important to both researchers and practitioners. Finally, the study addressed an important gap in the research literature by evaluating the factor structure of writing skills in a population about whom relatively little is known, young adults who have dropped out of high school.

20

References

Anderson, J.C. and Gerbing, D.W. 1984. The effect of sampling error on convergence, improper solutions, and goodness of fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49, 155-173.

Bentler, P.M. 1980. Multivariate analysis with latent variables: Causal modeling. In M.R. Rosenzweig and L.W. Porter (Eds)., Annual review of psychology, 31, 419-456.

Bentler, P.M. and Bonett, D.G. 1980. Significance tests and goodness of fit in the analysis of covariance structures. Psychological bulletin, 88, 588-606.

Birenbaum, M. and Tatsuoka, K.K. 1982. On the dimensionality of achievement test data. Journal of educational measurement, 19, 259-266.

Boomsma, A. 1982 The robustness of LISREL against small sample sizes in factor analysis models. In K.G. Joreskog & Wold (Eds.), Systems under indirect observation: Causality, structure, prediction (Part 1, pp. 149-173). Amsterdam: North-Holland.

Boomsma, A. 1985. Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. Psychometrika, 50, 229-242.

Campbell, D. and Fiske, D. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological bulletin, 56, 81-105.

Carmines, E. and McIver, J. 1981. Analyzing models with unobserved variables: Analysis of covariance structures. In G. Bohrenstedt and E. Borgatta (Eds.), Social measurement: Current issues. Beverly Hills:Sage.

General Educational Development Testing Service. 1984. GED Tests specifications committee report. Washington, D.C.:American Council on Education.

General Educational Development Testing Service. 1987. The official teacher's guide to the Tests of General Educational Development. Washington, D.C.:American Council on Education.

Gorsuch, R.L. 1983. Factor analysis (2nd ed.). Hillsdale, N.J.:Lawrence Erlbaum Associates.

Hayduk, L.A. 1987. Structural equation modeling with LISREL: Essentials and advances. Baltimore, Md.:John Hopkins University Press.

Hoelter, J.W. 1983. The analysis of covariance structures: Goodness-of-fit indices. Sociological methods and research, 11, 325-344.

James, L.R., Mulaik, S.A., and Brett, J.M. 1982. Causal analysis: Assumptions, models, and data. Beverly Hills:Sage.

Joreskog, K. 1979. Analyzing psychological data by structural analysis of covariance matrices. In K.G. Joreskog and D. Sorbom (Eds.), Advances in factor analysis and structural equation models. Cambridge, Mass.:Abt.

Joreskog, K.G., & Sorbom, D. 1983. LISREL: Analysis of linear structural relationships by the method of maximum likelihood, User's Guide, Versions V and VI (2nd ed.). Chicago: National Educational Resources, Inc.

Kavanagh, M.J., Mackinney, A.C., & Wolins, L. 1971. Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological bulletin, 75, 34-49.

Loehlin, J.C. 1987. Latent variable models: An introduction to factor, path, and structural analysis. Hillsdale, N.J.:Lawrence Erlbaum Associates.

Long, J. 1983. Confirmatory factor analysis. Quantitative applications in the social sciences. Beverly Hills:Sage.

Marsh, H.W. & Hocevar, D. 1983. Confirmatory factor analysis of multitrait-multimethod matrices. Journal of educational measurement, 20, 231-248.

Phillips, S.E. and Mehrens, W.A. 1987. Curricular differences and unidemensionality of achievement test data: An exploratory analysis. Journal of educational measurement, 24, 1-16.

Rock, D.A. and Werts, C.E. 1979. Construct validity of the SAT across populations -- An empirical confirmatory study. (Report No. RR-79-2). Princeton, N.J.:Educational Testing Service.

Stanley, J.C. 1961. Analysis of unreplicated three way classifications with application to rater bias and trait independence. Psychometrika, 26, 205-219.

Werts, C.E., Joreskog, K., and Linn, R. 1972. A multitrait-multimethod model for studying growth. Educational and psychological measurement, 32, 655-678.

Table 1. Goodness of Fit Statistics for CFA Trait and Method Models of Writing Skills of High School Seniors as Measured by the GED Writing Skills Test-Part I

Models	Chi-square	df	χ^2/df	GFI	RMR	NR>2.0	PFI
1. Null	20,617.70	105	196.35	.217	.504	all	
2. 1 Trait	621.54	90	6.91	.962	.023	10	.83
3. 2 Traits (1 & 2 combined)	617.86	89	6.94	.962	.023	10	.82
5. 2 Traits (2 & 3 combined)	620.65	89	6.97	.962	.023	10	.82
7. 2 Methods (1 & 2 combined)	617.10	89	6.93	.962	.023	10	.82

Table 2. Factor Structure Invariance Over Samples of Graduating High School Seniors and Young Adult GED Examinees: Summary of Fit Statistics

Hypothesis	N	Chi-square (df)	GFI	RMR	Chi-square /df	P-value
SIGMA	2532/699	397.89 (120)	.951	.146	3.32	.000
	400/400	139.70 (120)	.980	.093	1.16	.106
k	2532/699	774.08 (180)	.972	.026	4.30	.000
	400/400	185.18 (180)	.972	.026	1.03	.380
LAMBDA	2532/699	806.27 (194)	.967	.044	4.16	.000
	400/400	198.87 (194)	.969	.037	1.02	.390
THETA	2532/699	935.39 (209)	.951	.050	4.48	.000
	400/400	242.94 (209)	.963	.039	1.16	.054
PHI	2532/699	970.08 (210)	.948	.147	4.62	.000
	400/400	256.21 (210)	.961	.095	1.22	.016

Attachment A

Writing Skill/Item Type Categories of Items for
15 Miniscale Variables from the GED Writing Skills Test, Part I

Variable Name	Number of Multiple-Choice Items	Skill/Item Type Categories
Variable 01:	3	Sentence Structure/ Sentence Correction
Variable 02	3	Sentence Structure/ Sentence Correction
Variable 03	3	Sentence Structure/ Sentence Revision
Variable 04	3	Sentence Structure/ Sentence Revision
Variable 05	5	Sentence Structure/ Construction Shift
Variable 06	3	Usage/Sentence Correction
Variable 07	3	Usage/Sentence Correction
Variable 08	3	Usage/Sentence Correction
Variable 09	3	Usage/Sentence Revision
Variable 10	3	Usage/Sentence Revision
Variable 11	3	Usage/Construction Shift
Variable 12	4	Mechanics/Sentence Correction
Variable 13	3	Mechanics/Sentence Correction
Variable 14	3	Mechanics/Sentence Correction
Variable 15	5	Mechanics/Sentence Revision